DOCUMENT RESUME

ED 464 091 TM 033 771

AUTHOR Snodgrass, Donna; Salzman, James A.
TITLE Creating the Rosetta Stone: Deciphering the Language of Accountability To Improve Student Performance.
PUB DATE 2002-04-00
NOTE 35p.; Paper presented at the Annual Meeting of the American Association for Educational Research (New Orleans, LA, April 1-5, 2002).
PUB TYPE Reports - Descriptive (141) -- Speeches/Meeting Papers (150)
EDRS PRICE EDRS Price MF01 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS *Accountability; *Achievement Tests; Elementary Secondary Education; *Models; State Programs; *Test Interpretation; Testing Programs
IDENTIFIERS Ohio; Ohio Fourth Grade Proficiency Test; Ohio Off Grade Proficiency Tests; Ohio Sixth Grade Proficiency Test; Ohio Twelfth Grade Proficiency Test

ABSTRACT

The objectives of this paper are threefold. First, a model is proposed for unifying massive amounts of conceptual and numerical information flowing from the measures of the accountability movement in Ohio and the materials that are publicly available to educators. Second, this model is translated into useable forms of information that help teachers drive instructional practices in their classrooms. Third, the efficacy of this model meant to improve state-mandated proficiency scores at the district level is discussed. The paper attempts to integrate numerical, pictorial, graphical and narrative information about the Ohio Proficiency Tests in a way that provides the reader with a rudimentary model of an educational Rosetta Stone. This tool helps educators decipher the contents of the Ohio proficiency tests at a level complex enough so educators can identify basic and fundamental instructional needs of their students. Six appendixes contain sample items from the Ohio proficiency tests. (Contains 1 figure, 3 tables, and 14 references.) (Author/SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 464 091

Running Head: ROSETTA STONE

# Creating the Rosetta Stone: Deciphering the Language of Accountability to Improve Student Performance

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY

D. Snodgrass

2A TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC).

Donna Snodgrass, Ph.D.
South Euclid-Lyndhurst City Schools
5044 Mayfield Road
Lyndhurst, OH 44124
216-9-691-2063
snodgrass@sel.k12.oh.us

James A. Salzman, Ph.D.
Cleveland State University
Rhodes Tower, Room 1341
1860 East 22nd Street
Cleveland, OH 44114-4435
216-687-5048
j.salzman@csuohio.edu

Presented at the American Educational Research Association

Annual Meeting, April 1-5, 2002, New Orleans, LA

TM033771



BEST COPY AVAILABLE

Abstract

The objectives of this paper are threefold. First, a model is proposed for unifying massive amounts of conceptual and numerical information flowing from the measures of the accountability movement in Ohio and the materials that are publicly available to educators. Second, this model is translated into useable forms of information that help teachers drive instructional practices in their classrooms. Third, the efficacy of this model meant to improve state-mandated proficiency scores at the district level is discussed.

Ultimately, the authors attempt to integrate numerical, pictorial, graphical and narrative information about the Ohio Proficiency Tests in a way that provides the reader with a rudimentary model of an *educational Rosetta Stone*. This tool helps educators decipher the contents of the Ohio proficiency test at a level complex enough so educators can identify basic and fundamental instructional needs of their students.

# Creating the Rosetta Stone: Deciphering the Language of Accountability to Improve Student Performance

## Introduction

In 1799, soldiers in Napoleon's army stumbled upon a block of stone as they were doing engineering work at Fort Julien, near Rosetta, approximately 35 miles northeast of Alexandria. The inscriptions on the stone dated to 196 BC to the reign of Ptolemy V. The decree was written in two languages, Egyptian and Greek, with the Egyptian being written twice, once in hieroglyphics and once in an Egyptian cursive script which evolved from hieroglyphics. The assumption was that all three languages had the same message. Though both Egyptian writings were indecipherable at the time, it followed that, working from the Rosetta Stone, scholars of Greek could use it to translate Egyptian hieroglyphics. From the discovery, however, it still took over 20 years before Jean Francois Champollion was able to crack the code and decipher the messages on Egyptian antiquities.

In some ways, the authors suppose that we have a sense of the task Champollion undertook when he set out to uncover the meanings of the hieroglyphs on the Rosetta Stone. In educational circles, high-stakes tests have provided teachers and other stakeholders with their own set of "decrees" written in a language that is neither spoken nor read, and certainly little understood, by the vast majority of professionals affected by them. We also are fortunate to find ourselves a little ahead of Champollion; it's only taken us a little over ten years to wade through the morass which has been the high stakes testing in the state of Ohio and develop ways of helping teachers decipher this little understood language. Our intent in this paper is to share some of our discoveries as we attempted to help teachers make sense of high-stakes tests.

The objectives of this paper are three-fold. First, we will propose a model that has evolved for unifying massive amounts of conceptual and numerical information flowing from the measures of the accountability movement in Ohio and the materials that are publicly available to educators. Second, we will demonstrate how we have translated this model into useable forms of information that help teachers drive instructional practices in their classrooms. Third, we will demonstrate the efficacy of this model to improve state-mandated proficiency scores at the district level in one district that used this model.

## Theoretical Perspective

Forty-nine of the 50 states have content standards which drive instruction within each state. Most of those states have a high-stakes testing program which compels teachers and districts to spend a great deal of time and energy preparing students to meet the objectives of the test. Some (see Kohn, 2000; Popham, 2001; Sacks, 1999; Smith 1998) would say that this is to the detriment of children educationally and psychologically. Others (see Tucker & Codding, 1998) would argue that without standards-based assessment, there can be no accountability. This view is also reflected in the latest education proposal to come out of the White House (No Child Left Behind Act 2001). Regardless of which side of the debate one finds oneself, high-stakes assessment is a reality with which most schools nationwide are faced. And one in which educators play an important role, not necessarily as advocates pro or con, but as gatekeepers to determine that tests are being used accurately and fairly for all who are impacted by them (Newman & Benz, 1996).

Ohio's High-Stakes Accountability System

The situation in Ohio is similar. Schools are now issued a report card in the spring of the year in which they are placed in one of five categories, based on how many of 27 report card indicators the district obtains. The five categories are ranked, from lowest to highest, as follows: *Academic Emergency, Academic Watch, Continuous Improvement, Effective,* or *Excellent*. Of the 27 indicators currently on the district report card, 25 of the indicators are based on student performance on the previous year's 4th, 6th, 9th and 12th grade Ohio proficiency tests in *mathematics, science, writing, reading,* and *citizenship*. This will change in the near future since 2001 was the final year that the twelfth grade proficiency test was administered; thus this year is the last year that twelfth grade test results will be part of the report card. These "report cards" are sent home to parents and results are widely published in Ohio newspapers. Ohio educators are under constant pressure to raise or maintain high proficiency test scores.

Although the Ohio Proficiency Tests began in 1990, until recently Ohio educators were not provided with a number of tools that facilitate teachers' abilities to decipher the contents and structure of the proficiency tests in a way that promotes teaching and learning. Test results were reported for districts, buildings, classrooms and individual children, but the results did not allow teachers to locate essential information about the instructional needs of students. Student test results did allow politicians and the news media to publicly rank schools based on the numbers of children who performed at state proficiency standards. However, the test results that were provided were too limited and presented in a way that at best were confusing and at worst were misleading for instructional purposes.

The Rosetta Stone of Accountability: The Keys to Reading Proficiency Test Hieroglyphics

The discovery of the Rosetta Stone allowed linguists to uncover the language of Egyptian hieroglyphics and ultimately to decipher the messages on a variety of antiquities. Though the antiquities were known of for a long time, the key to decoding them was missing. For high stakes tests, there are several keys that teachers and other stakeholders need. As we have conceptualized them, at minimum, teachers need:

1. Specific content standards around which the test will be organized.
2. Performance standards and an adequate number and variety of specific concrete examples of the types of items on the tests.
3. Item analyses reports that reveal the proportion of students who are able to pass each item on the test.
4. Easy access to graphically organized student data.

The Need for Content Standards

For the past ten years the Ohio Proficiency Tests (OPT) have not been based on content standards. Content standards provide educators with an understanding of the fundamental knowledge and skills that are required of students taking the tests. They provide teachers with information about the cognitive capacities and essential thinking strategies that students must demonstrate to pass a test. They provide the schematic structure around which test items can be designed, curriculum can be developed and instructional practices can be organized. In their current state, the OPTs were built before the standards in most of the content areas. Add to this the fact that Ohio is revisiting their

standards, already adopting new English Language Arts (ELA) and mathematics standards in December 2001 (so that there will be new standards adopted while the old tests are being used) and one can imagine the uncertainty and anxiety facing Ohio's teachers.

The Need for Performance Standards and Concrete Examples of Test Items

Ohio's proficiency tests not only lack content standards, but there are no performance standards associated with the tests. Performance standards describe the types of behavior students must demonstrate to show mastery of the content standards. Performance standards highlight what tasks or behaviors the students must demonstrate and how well the students must be able to demonstrate them to pass the test. The situation was further complicated by the fact that until a court order in July 1998, teachers did not have available to them past test items, which would provide an adequate number and variety of sample test items to employ as reference guides.

For the vast majority of that time, Ohio teachers were provided with only "theoretical statements" about the contents of the tests, written in the form of abstract definitions of *strands and outcomes*. From these theoretical statements, teachers were asked to divine the particular problems the state intended to test. For example, educators were not provided with performance standards or exemplary items for the strand *Numbers and Number Relations*. Rather, they were told that fourth grade students should be able to "decompose, combine, order and compare numbers" (*Numbers and Number Relations*; Outcome #6). Teachers would have to play a guessing game with the state as to which specific items would be tested in the domain of all possible items that require a fourth grader to *decompose, combine, order* and *compare numbers*. Without content

standards, performance standards or a series of exemplary items this situation is, at best, a daunting task.

In their report, *Illustrative Language for an RFP to Build Tests to Support Instruction and Accountability*, The Commission on Instructionally Supportive Assessment (2001), chaired by W. James Popham, noted: "A state's high-priority content standards must be clearly and thoroughly described so that the knowledge and skills students need to demonstrate competence are evident" (p. 8). However, Ohio teachers were presented with neither clear and thorough content standards, nor performance standards. Moreover, for the majority of the past decade, Ohio educators were not provided with a variety of well-targeted sample items around which teachers could organize their instructional practices. Without these three elements teachers do not know what they should teach and children do not know what they should learn.

The Need for Item Analyses

For most of the past 11 years, educators also were not provided with any information about how students performed on individual test items. They were merely provided with composite indicators of how their students performed relative to general strands. Moreover, there were no differential indicators in proficiency reports of how students performed on multiple-choice versus constructed response items.

Item analyses reports were not available to Ohio educators until February 2000. Even when they were available, their existence was not well publicized. They could only be obtained from the scoring contractor at an additional cost to the school districts, and they were only available to school districts for a limited period of time.

Once they were in the hands of educators, the item analyses ultimately revealed the logical and statistical fallacy of providing educators with a single composite indicator of how students performed on strands, rather than providing educators with information about how students performed on individual items within strands. The reports showed the degree to which student mastery of different types of items varied within strands.

As an illustrative example, we present Table 1, which shows the percentage of district students who were able to pass each of the four items from the strand *Geometry* on the 2000 Fourth Grade Mathematics Proficiency.

Table 1

Percentage of District Fourth Graders Who Passed the Four Geometry Items on the 2000 Ohio *Fourth Grade Mathematics Proficiency Test* by Item Type and Item Point Value

| Items | Percentage Passing the Item | Type of Item | Points Value |
|---|---|---|---|
| 1. | 67% | Multiple-choice | 1 |
| 2. | 36% | Constructed-response | 2 |
| 3. | 95% | Multiple-choice | 1 |
| 4. | 67% | Multiple-choice | 1 |

Rather than reporting individual rates of student passage on the four items that measure the strand *Geometry*, the state collapsed the four passage rates (67%, 36%, 95%

and 67%) and reported them as single composite indicators on *Subscale* (strand) *Frequency Distributions*. These composite indicators masked the fact that students performed markedly differently on the different items within the strand *Geometry* (as well as most other strands). In addition, without item analysis reports, educators could not see the obvious difficulty students had correctly responding to constructed-response items. Not only did the students perform more poorly on the single constructed-response item in *Geometry* (38%), the item had the highest point value (2 points). The item analyses reports showed a clear pattern of lower scores among items requiring constructed-responses. Creating constructed-responses in *Geometry*, as well as other areas was a skill that needed to be addressed, but without a report that disaggregated this information, this instructional implication was hidden from teachers.

One of the cardinal rules of data management is that categorically collapsed data always wipes out detail and reduces the specificity of information. Only when teachers have detailed information about their students' performance can they assess the instructional needs of their students and evaluate the effectiveness of their teaching methods. The item analyses reports, tied to the newly released items, would for the first time provide educators with the beginnings of a Rosetta Stone, which would ultimately lead educators to more defensible instructional decisions.

Item analyses would not only highlight what skills needed to be taught, but these reports would also provide educators with a number of useful comparison groups. Using item analysis reports, educators could see the relative passage rates of items for students *across the state*, students *in their districts*, students *in their buildings*, and students *in their classrooms*. Meaningful data always exist in context rather than isolation, and these

comparison groups provided reference points, which allowed educators to better understand the relative effectiveness of their instruction, at least as it was measured on one high-stakes test.

The Need for Graphical Displays of Data

Aggregate and disaggregate data about student performance on the Ohio Proficiency Tests comes in the form of basic descriptive statistics on vendor printouts. Most of these data and their associated data analysis techniques are abstract and dependent upon various computations. Unfortunately, most educational stakeholders are not comfortable with numerical representations of test scores, nor familiar with statistical analyses. In order to communicate effectively with teachers, there is precedent to de-emphasize statistical representations of data in favor of graphical representations (Landwehr & Watkins,1986; Tufte, 1983). Landwehr and Watkins, for instance, cite the Quantitative Literacy Movement, based on Tukey's (1977) work, that data should be used to find "patterns and surprises" (preface).

For those who are responsible for reporting on these patterns and surprises, those who are responsible at the district or building levels for evaluating educational programming, it is important that their reports and resources are tailored to the audiences for whom they are meant (Worthen & Sanders, 1987). In schools, the greatest impact will be made when that information is able to inform both classroom teachers and building principals about student achievement and needs. In order to be heard by and to make a difference with those on whom they impact, Weiss and Bucuvalas (1980) noted that the dissemination of this information must meet two criteria. It must have both *truth value* and *utility value* for the recipients. In other words, results must be reported in such

a way that those who receive them recognize their own experiences and expectations in them, as well as seeing that they provide recipients with a direction that they can take to improve on the situation.

In order to begin to understand a multivariate and complex world, those employing visual displays to investigate the world around them must bring together multiple relevant factors (Tufte, 1983). Displays of proficiency test results must bring together factors like numbers, pictures, words, colors, and graphs in ways that allow teachers to develop meaningful and complex understandings of their own classroom realities. Well-organized visual displays of data can function as quantitative arrows that point the way to the qualitative needs of learners. With this as a foundation, for instance, one district has used graphical representations of data to provide early identification of at-risk students at different levels of need for intervention (Snodgrass & Salzman, 1998). This current study was designed to follow that district as it created visual resources to be used within the district to further improve students' performance on state proficiency tests.

## Creating Resources to Help Teachers Make Sense of Data

In order to unify the multiple sources of information that was swirling about the state, a group of district educators pulled these multiple sources into a set of visually-meaningful resources that became Proficiency Test Notebooks for teachers and principals. The purpose in creating these notebooks was to help teachers develop a more meaningful and complex understanding of the multiple variables that can impact their classroom reality. The notebooks were placed into the hands of teachers and unified the following information:

1. Multiple examples of past proficiency test items.
2. Information written beside each item that identified the strand and outcome the item is intended to measure.
3. State definitions of strands and outcomes.
4. Data about the proportion of students in the state, district and building that passed each item included in the notebook.
5. The correct response to each item.
6. A visual organizer that allows the user to quickly locate any item in the notebook based on any of the following criteria: strand, outcome, item type and more.
7. A list of content-specific and general vocabulary words from three years of past tests.

These notebooks were created with the intent to unify, organize and highlight proficiency test information in a way that provided educators with the keys for making instructional sense of the proficiency test. The *Fourth Grade Mathematics Notebook* was the first to be designed, followed by a *Fourth Grade Science Notebook*. This study evaluates how the mathematics and science proficiency notebooks played a role in helping to impact a district's fourth grade proficiency test scores.

Organization of the Teacher Resource Notebook

The *Fourth Grade Mathematics Notebook* includes over 100 items taken directly from three years of past fourth grade mathematics proficiency tests. The test items were cut apart, reorganized and compiled within the notebook by the 25 outcomes, which are organized under the 8 strands that make-up the *Ohio Fourth Grade Mathematics*

*Proficiency Tests*. These fourth grade mathematics strands and outcomes are listed in Table 2.

Table 2

The 8 Strands and 25 Outcomes that Are Measured on the Ohio Fourth Grade Proficiency Tests

| | Strands | Outcomes |
|---|---|---|
| 1. | Patterns Relations & Functions | 1, 2 |
| 2. | Problem Solving | 3, 4, 5 |
| 3. | Numbers & Number Relations | 6, 7, 8, 9, 10, 11 |
| 4. | Geometry | 12, 13, 14 |
| 5. | Algebra | 15, 16 |
| 6. | Measurement | 17, 18, 19, 20, 21 |
| 7. | Estimation & Mental Computations | 22, 23 |
| 8. | Data Analysis & Probability | 24, 25 |

Each test item is placed on a separate page in the notebook (See Appendix A for example). A number of other pertinent pieces of test information are placed on the page next to each item. First, the *strand name* and *outcome number* the item is intended to measure are placed on the page beside the item (The state's definitions of each strand and outcome are inserted in separate section at the beginning of the notebook). Second, the *proportion* of *Ohio students*, *district students* and *building students* who were able to pass the item is also

written on the page. Third, the correct answer to the item is identified. In the case of each multiple-choice item the correct response is placed directly on the item. In the case of a constructed-response item the state's scoring guideline for that item is located on the page(s) directly following the item.

The notebooks also include a visual organizer, which functions as a *table of contents* (see Appendix B). This visual organizer is located in the first section of the notebook. Using this organizer, teachers can quickly locate all items in the notebook by *strand, outcome, item type* (multiple-choice or constructed response), *date of administration* of the test item and more.

Like the notebook, the visual organizer is organized by strand and outcome, and can reveal unexpected surprises within the tests. A quick analysis of the visual organizer reveals interesting patterns in the mathematics proficiency tests that are not readily seen when looking through past test booklets themselves. For example, the visual organizer shows that 3 of the 6 items over the past three years that are the test's most challenging items, *extended response items*, come from the *Measurement* strands. By organizing numbers and words in this way, teachers can discover *patterns and surprises* that are not readily seen when the numbers and words are in their unorganized form.

The notebook also includes a section that lists *content-specific* and *general vocabulary words* from the past 3 years of the *Fourth Grade Mathematics OPT*. The vocabulary words make-up the last three pages of the book. Understanding and accurately employing these vocabulary words is essential to academic success in the content areas. Unlike literature-based reading, where readers can glean the meaning of the text even with a general or colloquial understanding of words, to succeed in content-specific, non-fiction

reading, learners must have explicit understanding of each technical word. Content area vocabulary words by their very natures have specific meanings and, to understand content area reading, students must have mastered and internalized those meanings. In this case, to succeed, both on the mathematics proficiency test and in mathematics in the world at large, students must employ mathematics vocabulary words in their most specific and literal sense.

Seemingly everyday words also can present a problem for students engaging in content area reading, when these common words form phrases that are unclear or unfamiliar to the students. For example, a child may have a problem understanding a test item when the child is accustomed to hearing his teacher use the phrase "each week," but the test employs the phrase "per week." The vocabulary list in the notebook provided a ready sources of technical and non-technical vocabulary words from the *Mathematics OPTs*, which could be systematically woven into daily lessons.

## District Procedures for Employing the Notebooks

The mathematics and science notebooks, their contents and their potential usage were presented to the district's six elementary principals by the notebooks' authors at an administrative retreat prior to the opening of the 2000/2001 school year. Principals were asked to share the notebooks and the information about the notebooks with their fourth grade teachers. The two notebooks were in the hands of all the district's fourth grade teachers in the fall of the 2000/2001 school year.

The notebooks were intended to be a tool for teachers to use and were employed in a variety of ways within the district. For example, one principal suggested to his teachers that they review only those multiple-choice items and their related concepts on which 50% or more of the district students were able to pass. After careful analysis of the

available information from reviewing the notebook items, he hypothesized that multiple-choice items with less than a 50% district passage rate were either poorly constructed or developmentally inappropriate. In either case, he chose not to have his teachers waste valuable instructional time on these types of concerns. Other principals empowered their teachers to determine to how best to use the information provided in the notebook. Though there was no one district-wide model for rolling out the information contained in the notebooks, all of the district fourth grade teachers used both the mathematics and science notebooks during the 2000/2001 school year.

These two notebooks provided teachers with enough exemplars of each strand and outcome that teachers could develop a conceptual understanding of how the state operationally defined the strands and outcomes. Providing teachers with the proportion of student passage for each item allowed them to identify the specific items that their students had mastered and target those that their students had failed to master. This information led teachers to more defensible decisions about the use of instructional time.

A principal of one of the district's traditionally lowest scoring elementary buildings commented to one of the authors of this paper that her fourth grade teachers told her that it was the sample items from the notebooks that pushed their students "over the top" on the mathematics proficiency test.

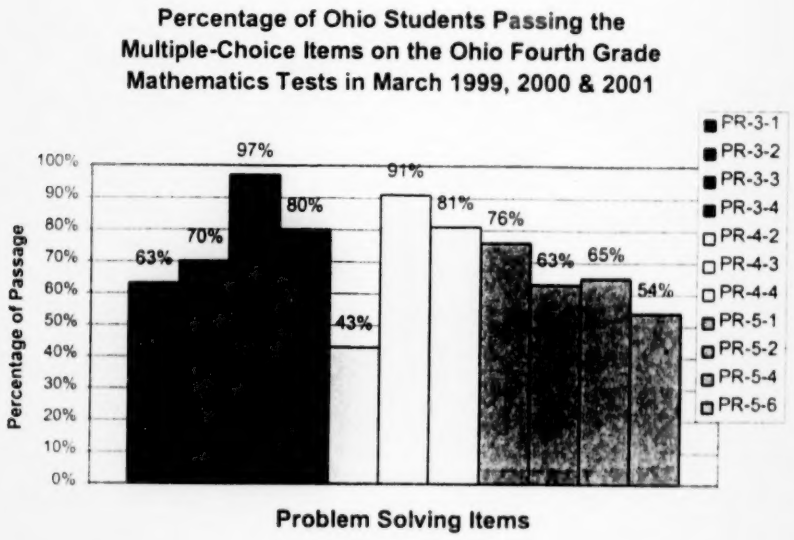
Visually Presenting Data to Help Teachers Decipher the Tests

Eight sets of bar graphs were later produced as companion tools to the *Fourth Grade Mathematics Notebook*, each representing one of the 8 different strands measured on the Fourth Grade Mathematics OPT (See Table 2 for a list of fourth grade

mathematics strands). Figure 1 shows the bar graph that represented the items from the ***Problem Solving*** strand.

---

Figure 1




---

Each bar on Figure 1 represents a single multiple-choice item from the Fourth Grade Mathematics Notebook. The bars are organized in the same order as the items are organized in the *Fourth Grade Mathematics Notebook*. The height of the individual bars illustrates the proportion of Ohio students who passed each multiple-choice item from the strand entitled *Problem Solving*. Individual bars within a given strand are grouped and shaded by outcome. The first four darkly shaded bars represent student performance on outcome #3. The next three lightly shaded bars represent student performance on outcome #4 and the final four bars represent student performance on outcome #5.

The letters and numbers in the graph's legend are page numbers from the *Fourth Grade Mathematics Notebook* and relate to the bar's corresponding notebook item. The legend reads as follows: **strand-outcome-sample item**. Therefore, **PR-3-1** reads as *Problem Solving Strand-Outcome #3-Sample Item 1*; **PR-3-2** reads as *Problem Solving Strand-Outcome #3-Sample item 2*, etc.

An analysis of the bars in Figure 1 highlight how visual displays of data can be used to quickly identify patterns that can lead to meaningful instructional practices. For example, the three lightly shaded bars in the center of Figure 3 representing multiple-choice items are, in theory, designed to measure the same outcome (#4). However, there is a noticeable variation in the proportion of Ohio students -- 43%, 91% and 81, respectfully -- who are able to pass each of the three items. Though ostensibly designed to measure the same domain of knowledge, the differences in the proportion of students able to pass the items lead one to question that assumption. Why, one might ask, is one item obviously so much more difficult for students to answer (43% rate of passage) than another item (91% passage rate)? These two test items are presented in appendices C and D as they appear in the proficiency test. An analysis of the actual two items reveals a surprise that might otherwise have remained hidden, providing teachers with instructionally useful information about the similarities and differences between the items.

A Closer Look at Surprises Revealed

Both items are short, story problems that require that students recognize information necessary to solve the problem. A quick review of the two items highlights differences between the two items. For instance, the "easier" item (Appendix C)

presented students with familiar concepts, like purchasing school supplies (pencils, notebooks, etc.) and receiving change. The more difficult item (Appendix D) also presented students with a short, story problem, but this time about hourly wages and the amount of money earned per week. *Hourly wages* is a concept that is certainly less familiar to fourth graders than purchasing school supplies and receiving change. Moreover, the easier item has a single, correct answer. The harder item requires that a child make a choice of which answer is the *best answer*. The typical fourth grader generally does not consider that there might be a *best answer*. Many fourth graders will likely choose the first answer that seems correct, even if that answer is only partially correct. Both A and B, for instance, are pieces of information that can help you solve this problem, though C is the only answer that provides all the information that students need. This is further complicated by the fact that students were never explicitly told within the prompt that there is, or that they should be looking for, a *best* answer. Students must infer that information from the prompt.

In this case, the differences in the difficulty of the items may be less related to the mathematical thinking the child is asked to perform than the story that the problem is embedded in and the structure of the item itself. The discrepancy between the scores on the two items (43% rate of passage versus 91% rate of passage) is a good example of how noticing a *quantitative* difference between the passage rates on two items can point to *qualitatively* different instructions need of children.

Engaging teachers in detailed analysis of items in this way leads to another instructionally useful situation. Two different test items, presented in Appendices E and F, are categorized under the strand *Numbers and Number Relations*. These items present

students with virtually the same task. Curiously, however, there is an 11% difference in the proportion of district students who correctly answered one item over the other item. Both items require that students identify the correct symbol "<, > or =" that shows the relationship between two simple fractions with unlike denominators. Further review of the two items shows that the item with the higher rate of passage, 72%, (Appendix E) includes two pie graphs. The more difficult item does not include the pie graphs. Students are instructed in the easier problem that they can use the pie graphs to help them solve the problem. The pie graphs, it appears, are designed to function as visual aids for the students. As the difference in the percentage of students who were able to pass the two items suggests, they do appear to assist the students. The instructional implications resulting from the differential rate of passage of these two problems are that visual aids, like pie graphs, assist students in solving problem or checking their work. Again, the *quantitative* differences in the magnitude of the two bar graphs help teachers to identify the *qualitative* instructional needs of learners.

Professional development in the district was built on these types of analyses of individual and groups of items purporting to represent the same strands and outcomes. Ultimately, creating the notebooks and graphs, and engaging educators in using them, resulted in their asking questions like:

- *Why do two items which purport to measure the same strand and outcome result in large differences in the percentage of students able to pass each of the items?*

- *Is there a relationship between the proportion of fourth grade students who pass an item and the developmental appropriateness of the item, the quality of the item or the way we teach the item?*
- *How does our students' performance compare to student performance across the state on each item?*
- *Do our children have the same level of mastery on constructed-responses as multiple-choice items?*

The notebooks and graphs cannot in and of themselves change student performance. Nor can the notebooks tell educators which items on the proficiency tests are developmentally appropriate and critical to students' understanding of mathematics and which items are not. These decisions are judgments. Good judgments do not exist in a vacuum or by playing a guessing game with the state. The graphs and notebooks can provide a way to decipher the contents of the test and help teachers make defensible instructional decisions. Only when resources, like these notebooks, are available and used as part of a systematic and integrated curricular initiative, will district personnel, both teachers and principals, maximize teachers' abilities to help students reach their potential as measured on high stakes tests.

## Results

The data in Table 3 show that there was an increase of 12% in the proportion of students able to pass the mathematics proficiency test in March 2001, when teachers employed the mathematics notebook, over the March 2000 administration of the test, when the notebook was not employed. Table 3 also shows a parallel increase in science

(+8%) from 2000 to 2001. Chi-square tests indicated that these gains were statistically significant.

Table 3

Comparison of the Percentage of District Students Who Passed the Ohio Fourth Grade Mathematics and Science Proficiency Tests between March 2000 and March 2001

| Test | March 2000 N | March 2000 % Pass | March 2001 N | March 2001 % Pass | % Diff |
|---|---|---|---|---|---|
| 1. Mathematics | 284 | 65% | 256 | 77% | +12%* |
| 2. Science | 285 | 59% | 255 | 67% | +8%* |
| 3. Writing | 283 | 88% | 250 | 90% | +2% |
| 4. Reading | 281 | 68% | 250 | 68% | 0% |
| 5. Citizenship | 285 | 75% | 249 | 76% | +1% |

Note. * indicates chi-square test significant at $p<.05$

The question of whether or not the increased proportion of student passing the Mathematics and Science tests might be attributed to factors other than the notebooks exists. For example, the improvement in the scores could be the result of an inherent difference in the ability of the two sets of students who participated in the March 2000 and March 2001 tests. One way to investigate the likelihood that the difference in the scores is a function of an inherent difference in the abilities of the two sets of students is to compare the two sets of fourth graders' scores on the other 3 proficiency tests: *Writing*, *Reading* and *Citizenship*, since notebooks were neither used nor developed in those areas. The percentage of passage in *writing* between March 2000 and March 2001 increased by

two percent (+2%). In *reading* the percentage of passage remained exactly the same (0% difference) both years (68%). In *citizenship* the percentage of passage increased by 1%. None of the gains were statistically significant. While there are clearly other plausible reasons for the differences in scores, it appears that the notebooks and adjunct materials may provide a feasible explanation for the improvement in mathematics and science proficiency scores between March 2000 and March 2001and not merely reflect any inherent difference in ability of the two different sets of fourth grade students.

## Educational Importance

The development of the workbooks seems to have had a three-fold benefit. First, large amounts of conceptual and numerical information, flowing from the measures of the accountability movement have been unified in a way that makes the intended outcomes of the *Ohio Fourth Grade Proficiency Test* more accessible to the major stakeholders. Second, this model has unified numbers, words and graphical information in ways that help teachers drive instructional practice in their classrooms. Finally, the efficacy of this model appears to be supported. A statistically- and practically-significant increased proportion of students did pass the mathematics and science tests during the school year when the notebooks were employed (2000/01) as compared to the proportion of students passing the mathematics and science tests during the school year prior to their development (1999/2000). In areas where notebooks did not exist (writing, reading and citizenship) increases in performance were not evident at a district level.

Based on these experiences, we propose that in order for statewide testing programs to be useful for instructional purposes educators must be provided with at least the following information:

1. Specific content and performance standards around which items are developed. Currently there are no apparent content or performance standards around which the Ohio Proficiency Tests are built. Though that may change with the development of new content standards in 2002, the state's current practices do not inspire confidence.
2. An adequate number and variety of specific concrete examples of the types of items on the proficiency test.
3. Item analyses reports that present the rate of student passage of each ***item*** by state, district, building and classroom results. Item analyses reports should be a standard contracted part of the test package negotiated by the state with the test vendor and, therefore, free and readily available to all districts. While there is great value for teachers and other district personnel to engage in the inquiry and analysis necessary to create resources like these, it is our belief that, if a state is going to hold schools accountable for meeting high-stakes accountability measures, they have a concomitant responsibility to invest the necessary monies to ensure that teachers are provided with the tools to help their students succeed.
4. Easy access to graphically organized student data presented in a form that is readily understandable to the statistically less inclined user.

## A Caveat

Currently in Ohio, rather than creating new test items after the old test items have been released to the public, many test items from all the tests (reading, writing, mathematics, science and citizenship) are being reused by the state in subsequent versions of the proficiency tests. This is arguably a questionable decision. For, there is little doubt that the reuse of test items after they have been released contributes to an

increase in the percentage of students who are able to meet Ohio standards on the Ohio Proficiency Tests. Moreover, the more savvy educator immediately realizes that as items are being released and reused less and less can be inferred from student test results about students' ability to perform items in the greater domain that the test purports to be measuring. This is an unfortunate situation.

The intent of the authors of the notebooks was not to provide items to teachers so teachers could teach *the test*. Rather, the purpose of the authors was to provide a rudimentary model of an *educational Rosetta Stone* that would help teachers decipher the contents of the proficiency test at a deep and complex enough level to identify the basic and fundamental instructional needs of students. These needs are based on skills and knowledge that always exist independent of any test that attempts to measure them.

## Epilogue

The intention of this paper was to introduce readers to the work done in trying to make sense of high-stakes testing information. The authors are part of a grassroots network (Support Team for Assessment, Testing, and Statistics or STATS) of testing supervisors, curriculum directors, university professors and other interested educators who meet monthly to discuss testing and assessment issues. Since these notebooks were created, other STATS-member districts, having heard about them and seen them, have used the template we've described and created other notebooks to address aspects of the test that had yet to be organized in this way. Recently, all of these notebooks were compiled, burned onto a CD and sent to every school district in the state of Ohio.

# Appendix A

## Fourth Grade Mathematics Proficiency Test

Test Date: 99
Item #: 31
Strand: PR
Outcome: 4
Item Type: (MC) SA ER

**Percentage of Student Passing**

State 43 %  District _____%
School _____%  School _____%
School _____%  School _____%

31. Mavis works at the hardware store. Her hourly wage is $4.50. How much money is Mavis paid for one week's work?

Which piece of information will help you solve this problem?

- ○ A. The number of hours she works each day
- ○ B. The number of days she works each week
- ☒ C. The number of hours she works each week

BEST COPY AVAILABLE

PR-4-2

Item Types and Point Value

| | |
|---|---|
| MC = Multiple-Choice | (1 point each) |
| SA = Short Answer | (2 points each) |
| ER = Extended Response | (4 points each) |

# Revised Fourth Grade Mathematics Proficiency Notebook

Table of Contents

**Section I**

State definitions of *Strands*, *Objectives* and *Levels of Understanding* in mathematics Pages 23 - 34

**Section II**

Examples of Strands, Objectives and Levels of Understanding taken from actual 1999 and 2000 Fourth Grade Proficiency Tests, as follows:

| Strand | Objective | Example | Type | Levels of Understanding | Strand/ Objective/ Example | Year/ Item No. |
|---|---|---|---|---|---|---|
| 1. Patterns, Relations and Functions | 1 | 1 | MC | C - Conceptual Understanding | PA-1-1 | 99-6 |
| 1. Patterns, Relations and Functions | 1 | 2 | MC | A - Application and Problem Solving | PA-1-2 | 00-2 |
| 1. Patterns, Relations and Functions | 1 | 3 | SA | C - Conceptual Understanding | PA-1-3 | 00-32 |
| 1. Patterns, Relations and Functions | 2 | 1 | MC | C - Conceptual Understanding | PA-2-1 | 99-10 |
| 1. Patterns, Relations and Functions | 2 | 2 | SA | C - Conceptual Understanding | PA-2-2 | 99-15 |
| 1. Patterns, Relations and Functions | 2 | 3 | MC | C - Conceptual Understanding | PA-2-3 | 00-26 |
| 1. Patterns, Relations and Functions | 2 | 4 | ER | C - Conceptual Understanding | PA-2-4 | 00-29 |
| 1. Patterns, Relations and Functions | 2 | 5 | MC | C - Conceptual Understanding | PA-2-5 | 01-10 |
| 1. Patterns, Relations and Functions | 2 | 6 | SA | C - Conceptual Understanding | PA-2-6 | 01-15 |
| 2. Problem Solving Strategies | 3 | 1 | MC | A - Application and Problem Solving | PR-3-1 | 99-35 |
| 2. Problem Solving Strategies | 3 | 2 | MC | A - Application and Problem Solving | PR-3-2 | 00-19 |
| 2. Problem Solving Strategies | 3 | 3 | MC | C - Conceptual Understanding | PR-3-3 | 00-22 |
| 2. Problem Solving Strategies | 3 | 4 | MC | A - Application and Problem Solving | PR-3-4 | 01-9 |
| 2. Problem Solving Strategies | 4 | 1 | ER | A - Application and Problem Solving | PR-4-1 | 99-11 |
| 2. Problem Solving Strategies | 4 | 2 | MC | A - Application and Problem Solving | PR-4-2 | 99-31 |

# Appendix C

24. At the school store, José bought 2 pencils for 10¢ each, a notebook for 65¢, and a candy bar for 40¢. To find out how much change he will get, you need to know

O A. how much 2 notebooks would cost.

X B. how much money he gave the salesperson.

O C. how much he saved by buying one notebook.

# Appendix D

31. Mavis works at the hardware store. Her hourly wage is $4.50. How much money is Mavis paid for one week's work?

Which piece of information will help you solve this problem?

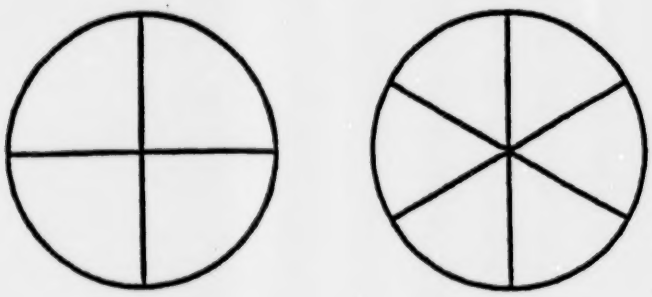
O A. The number of hours she works each day

O B. The number of days she works each week

X C. The number of hours she works each week

BEST COPY AVAILABLE

# Appendix E



Compare the fractions. Then choose the correct symbol.

You may use the pictures to help you solve the problem.

$$\frac{2}{4} \square \frac{3}{6}$$

- ⊗ A. =
- ○ B. <
- ○ C. >

# Appendix F

**32. Compare the two fractions. Then choose the correct symbol.**

$$\frac{2}{3} \square \frac{1}{2}$$

- (X) A. >
- ( ) B. <
- ( ) C. =

Works Cited

Kohn, A. (2000). The case against standardized testing: Raising the scores, ruining the schools. Portsmouth, NH: Heinemann.

Landwehr, J. M., & Watkins, A. E. (1986). Exploring data. Palo Alto, CA: Dale Seymour Publications.

Newman, I., & Benz, C. (1996 October). Standardized Testing. Paper presented at the Mid-Western Educational Research Association, Chicago, IL.

Ohio Department of Education. (1994). Language Arts: Ohio's Model Competency-Based Program. Columbus, OH: Author.

Popham, W. J. (2001). The truth about testing: An educator's call to action. Alexandria, VA: Association for Supervision and Curriculum Development.

Commission on Instructional Support Assessment (2001, October). Illustrative language for an RFP to build tests to support instruction and accountability. Chair: Popham, W. J.

Sacks, P. (1999). Standardized minds: The high price of America's testing culture and what we can do to change it. Cambridge, MA: Perseus.

Smith, F. (1998). The book of learning and forgetting. New York: Teachers College Press.

Snodgrass, D. G., & Salzman, J. A. (1998). Graphical representations of data: Iowa Tests of Basic Skills, Ohio sixth grade proficiency tests, and SB 55. Paper presented at the Mid-Western Educational Research Association, Chicago, IL. (ERIC Document Reproduction Services No.

Tucker, M. S., & Codding, J. B. (1998). Standards for our schools: How to set them, measure them, and reach them. San Francisco: Jossey-Bass.

Tufte, E. R. (1983). Visual display of quantitative information. Cheshire, CT: Graphics Press.

Tukey, J. W. (1977). Exploratory data analysis. Reading, MA: Addison Wesley.

Weiss, C. H., & Bucuvalas, M. J. (1980). Tests and utility tests: Decision makers' frames of reference for social science research. American Sociological Review, 45, 302-313.

Worthen, B. R., & Sanders, J. R. (1987). Educational evaluation: Alternative approaches and practical guidelines. New York: Longman.

TM033771

**U.S. Department of Education**
*Office of Educational Research and Improvement (OERI)*
*National Library of Education (NLE)*
*Educational Resources Information Center (ERIC)*

ERIC

# REPRODUCTION RELEASE

(Specific Document)

## I. DOCUMENT IDENTIFICATION:

Title: Creating the Rosetta STONE: Deciphering the LANGUAGE of Accountability to Improve STUDENT PERFORMANCE

Author(s): DONNA SNODGRASS & James A. SALZMAN

Corporate Source:

Publication Date: April 2, 2002

## II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

Level 1 ☐

Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) *and* paper copy.

The sample sticker shown below will be affixed to all Level 2A documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2A

Level 2A ☑

Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only

The sample sticker shown below will be affixed to all Level 2B documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2B

Level 2B ☐

Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

*I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.*

**Sign here, → please**

| | |
|---|---|
| Signature: Donna Snodgrass | Printed Name/Position/Title: TEST Supervisor |
| Organization/Address: South Euclid-Lyndhurst BOE<br>5044 Mayfield Rd.<br>Lyndhurst, Ohio 44124 | Telephone: 216-691-2063 FAX: 216-691-2294<br>E-Mail Address: SNODGRASS@SELO.K12.OH.US Date: 4-2-02 |

(over)

## III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, *or*, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

| |
|---|
| Publisher/Distributor: |
| Address: |
| Price: |

## IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

| |
|---|
| Name: |
| Address: |

## V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

University of Maryland
ERIC Clearinghouse on Assessment and Evaluation
1129 Shriver Laboratory
College Park, MD 20742
Attn: Acquisitions

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document bein contributed) to:

**ERIC Processing and Reference Facility**
**4483-A Forbes Boulevard**
**Lanham, Maryland 20706**

**Telephone: 301-552-4200**
**Toll Free: 800-799-3742**
**FAX: 301-552-4700**
**e-mail: ericfac@inet.ed.gov**
**WWW: http://ericfac.piccard.csc.com**

EFF-088 (Rev. 2/2000)